A Low-complexity Psychometric Curve-fitting Approach for the Objective Quality Assessment of Streamed Game Videos

Sam Van Damme, Maria Torres Vega, Joris Heyse, Femke De Backere and Filip De Turck

IDLab, Department of Information Technology, Ghent University - imec

Abstract

The increasing popularity of video gaming competitions, the so called *eSports*, has contributed to the rise of a new type of end-user: the passive game video streaming (GVS) user. This user acts as a passive spectator of the gameplay rather than actively interacting with the content. This content, which is streamed over the Internet, can suffer from disturbing network and encoding impairments. Therefore, assessing the user's perceived quality, *i.e.* the Quality of Experience (QoE), in real-time becomes fundamental. For the case of natural video content, several approaches already exist that tackle the client-side real-time QoE evaluation. The intrinsically different expectations of the passive GVS user, however, call for new real-time quality models for these streaming services. Therefore, this paper presents a real-time Reduced-Reference (RR) quality assessment framework based on a low-complexity psychometric curve-fitting approach. The proposed solution selects the most relevant, low-complexity objective feature. Afterwards, the relationship between this feature and the ground-truth quality is modelled based on the psychometric perception of the human visual system (HVS). This approach is validated on a publicly available dataset of streamed game videos and is benchmarked against both subjective scores and objective models. As a side contribution, a thorough accuracy analysis of existing Objective Video Quality Metrics (OVQMs) applied to passive GVS is provided. Furthermore, this analysis has led to interesting insights on the accuracy of low-complexity client-based metrics as well as to the creation of a new Full-Reference (FR) objective metric for GVS, *i.e.* the Game Video Streaming Quality Metric (GVSQM).

Keywords: Game video streaming (GVS), Quality of Experience (QoE), predictive modelling, objective quality assessment, curve-fitting, Game Video Streaming Quality Metric (GVSQM)

1. Introduction

The increasing interest in video gaming competitions (the so-called *eSports*) has risen the growth of the passive game video streaming (GVS) community, in which the end user is only watching the gameplay provided by other players instead of interactively participating [1]. This community has become so large that game-related live-stream web services have rapidly gained in popularity [2]. The most well-known example of such a platform is *Twitch*, with over 100.000.000 of unique users on a monthly basis [3]. These streaming services involve, however, network and video encoding related impairments such as delay, packet loss, jerkiness, frame rate or bitrate [4] that could negatively influence the user's perception of the service, *i.e.* the Quality of Experience (QoE) [5]. In order for internet-based GVS providers to be competitive with each other, it is thus of great interest to assess the perceived quality of the end-user in real-time. Based on these observations, multiple parameters, such as capacity and latency, could be adapted to maximize the end-user quality within the constraints of the platform, such that customers can be kept as satisfied as possible [6].

Given the humanly essence of perception, QoE has been traditionally measured by means of subjective experiments. During these, human observers score the content, leading to the so-called Mean Opinion Scores (MOS) [7]. Although being very accurate, these subjective tests are most often performed in limited laboratory environments and have high costs in terms of time, money and effort. In addition, MOS cannot straightforwardly be applied for real-time quality evaluation [7], which is essential for the real-time assessment in online streaming environments. Although continuous subjective evaluation methods exist (*i.e.* Single Stimulus Continuous Quality Evaluation (SSCQE), Double Stimulus Continuous Quality Evaluation (DSCQE)...)[8], these methods tend to be too intrusive for the end-user. As a matter of fact, they require the adaptation of a slider (or similar) for the real-

Email address: {firstname.lastname}@ugent.be (Sam Van Damme, Maria Torres Vega, Joris Heyse, Femke De Backere and Filip De Turck)

Pre-printed version. Please cite as: S. Van Damme, M. Torres Vega, J. Heyse, F. De Backere and F. De Turck, "A Low-complexity Psychometric Curve-fitting Approach for the Objective Quality Assessment of Streamed Game Videos," *Signal Processing: Image Communication*, vol. 88, no. 115954, 2020, doi: 10.1016/j.image.2020.115954

time indication of the perceived image quality. Especially in the case of interactive gaming, this is a problem as it prohibits the user to interact naturally with the game at hand. Even in the passive case, however, continuous evaluation methods distract the user from fully focusing on the provided gameplay. As such, subjective evaluations are not well suited for real-time quality assessment, which is an important condition in order to maximize the perceived quality of the end-user.

Alternatively, Objective Video Quality Metrics (OVQMs) are often used to model the quality of natural videos. These metrics aim at mathematically describing the subjective, human perception of visual quality by end-users as closely as possible. This is done based on numeric, visual characteristics of the received content as well as objective system factors, *e.g.* related to the network. According to their requirements, three types of models can be distinguished: *Full-Reference (FR)*, *Reduced-Reference (RR)* and *No-Reference (NR)*-models [5, 9]. FR-models calculate a quality metric by means of a mathematical comparison between the original and the distorted video. These approaches are however impractical in client-server based scenarios, as simultaneous access of both the original and distorted content is required [5]. The former is inaccessible at the client-side, however, due to the inevitable distortions that come with compression and network transmission. NR-models, on the other hand, attempt to estimate the quality of the content only based on the distorted video, at the cost of lower correlation with subjective scores but ruling out the problem of simultaneous access [5]. RR-metrics hold the middle between both as the original, undistorted content is compressed by calculating a number of features (*i.e.* low-complexity NR measurements) which can be sent over a side-channel to the client (using a limited part of the available bandwidth). This allows comparison with the distorted content to make a quality estimation [5]. Their correlation with MOS is usually better than their NR counterparts, but still significantly lower than FR solutions.

To provide a solution to the conflicting requirements of accuracy and computational complexity, either subjective scores or objective FR metrics are typically predicted from both encoding and/or network-related NR/RR metrics in order to make real-time quality estimations [5, 10–12]. Most of these approaches are applied to the case of natural videos, *i.e.* non-synthetic videos with real-life content, real actors etc. [13], often resulting in rather complex and computationally expensive relationships between the input features, *i.e.* the Quality of Service (QoS), and QoE [10, 11, 14, 15]. However, the expectations of passive GVS are intrinsically different (*e.g.*, more attention to moving objects, different perception of synthetic content, higher sensibility to fluidity...). As a result, it is important to understand whether the modelling approaches and accuracy levels of objective metrics for natural videos still hold for the passive GVS case. If this is the case, it is of great interest to investigate whether the typical characteristics of game video content can be exploited to obtain more straightforward, computationally friendly quality assessment methods than the often complex models applied for natural videos. If this is not possible, a need arises for alternative models focusing on the passive GVS end-user.

The goal of this work is therefore to provide an end-to-end solution for the real-time objective quality assessment of passive GVS. This solution consists of a RR framework based on a psychometric curve-fitting approach running on low-complexity, client-based objective metrics. As the most relevant metric will heavily depend on the video type [16–18], the presented end-to-end model includes a server-based pre-processing method. In this method, the best suited low-complexity metric (in terms of accuracy) is selected for each GVS sequence type being offered. In this paper, the working principles of this approach are presented. To illustrate its performance, the approach is applied to the *Gaming VideoSET* [19], a large dataset including subjectively scored sequences of passive GVS. To provide the accuracy analysis (the server-based method), 19 NR/RR features are selected, both on the pixel and bitstream level. In addition, 4 objective FR metrics are also put to the test with regard to the evaluation of their accuracy to subjective perception. As an output to the analysis, a game classifier is proposed and a customized FR metric for GVS, *i.e.* the Game Video Streaming Quality Metric (GVSQM), is created. Finally, the psychometric curve fitting approach is created and evaluated by benchmarking it (*i.e.* comparing to a certain subjective or objective FR ground truth) against both objective state-of-the-art FR metrics and subjective MOS. Furthermore, its accuracy is compared to two Machine Learning (ML) approaches often applied to natural videos: Decision Regression Trees (DRTs) and Artificial Neural Networks (ANNs).

The remainder of this paper is organized as follows. In Section 2, the related work within the field of predictive QoE modelling for GVS is presented and the shortcomings of the existing approaches are addressed. Section 3 provides a high-level overview of the theoretical approach followed. Next, a description of the used dataset is provided in Section 4. Section 5 highlights the most important results of the correlation analysis on a set of objective metrics often used for natural video quality estimation. The resulting game classifier as well as a customized, objective FR metric (GVSQM) are discussed as well. In Section 6, the proposed psychometric curve-fitting approach is further explained and evaluated and a comparison with other modelling approaches often applied in literature is provided. In Section 7, a short discussion is given on the pros and cons of the proposed approach in comparison with existing solutions. Section8, at last, provides a summary of the most important findings of this work, together with a few future research directions that could prove to be interesting extensions to the obtained results.

Table 1: Overview of the related work. The interaction type (*Int.*) is indicated with I for interactive and P for passive. *Type* indicates the nature of the metric presented in the study. *None* indicates that only a subjective study was performed, without proposing an objective metric. In addition, the considered distortions are indicated.

Authors	**Int.**	**Type**	**Distortions**
Clincy et al. [20]	I	None	• Latency • Packet loss
Huang et al. [21]	I	None	• Bitrate • Framerate • Resolution • Delay
Jarschel et al. [22]	I	None	• Delay • Packet loss
Slivar et al. (2015) [16]	I	NR	• Bitrate • Framerate
Slivar et al. (2016) [17]	I	NR	• Bitrate • Framerate
Slivar et al. (2018) [18]	I	None	• Bitrate • Framerate
Wang et al. [23]	I	NR	• Bitrate • Framerate • Resolution • Compression • Bandwidth limitation • Delay • Jitter • Packet loss
Zadtootaghaj et al. [24]	I	FR	• Bitrate • Framerate
Barman et al. (2018) [25]	P	None	Compression
Barman et al. (2018) [26]	P	NR	Compression
Barman et al. (2019) [27]	P	NR	Compression
Göring et al. [28]	P	NR	Compression

2. Related work

Only a limited number of studies exists up till now within the research field of visual quality modelling for GVS. An overview of these studies is provided in Table 1. As indicated in the previous Section , a distinction can be made between interactive and passive GVS.

Clincy et al. [20] present a study on the influence of network distortions on the client-side QoE in interactive GVS. The graphical part of the QoE is modelled in terms of packet loss while the interaction is measured in terms of the network latency. Their results show that both have a significant impact on QoE, but that the impact of packet loss is higher than the impact of latency. Moreover, the user tolerance to network distortions is shown to be lower for high-paced games such as First-Person Shooters (FPSs) in comparison with slower-paced games such as role-playing games (RPGs).

Huang et al. [21] perform a correlation analysis on the encoding bitrate, framerate and resolution at the encoding level combined with the network delay as a measurement for the interaction quality in an interactive GVS scenario. Their main conclusion is that resolution has a lower impact on the end-user QoE in comparison with the other measurements.

Jarschel et al. [22] show that the influence of the game type is less pronounced for network-related distortions in interactive GVS. Their correlation analysis shows that both network delay and packet loss have an important impact on the end-user QoE in terms of MOS, whereas the game genre and the skill of the particular player seem negligible.

Slivar et al. [16] attempt to model the visual QoE of interactive GVS purely on bitrate and framerate. These metrics are related to MOS using a linear regression method. The results show a heavy dependency on the game genre, with (rather limited) correlations ranging from 0.676 to 0.808 from game to game. In a second study [17], they did an attempt to improve their results by including a spatio-temporal characterization of the game as well as additional encoding characteristics. These features were related to MOS using a polynomial regression approach, resulting in increased correlations ranging from 0.782 to 0.986, but still heavily depending on the type of game. In a successive, third study [18], research is performed towards the most optimal encoding strategy given a particular game genre and some player-related characteristics, *e.g.* the experience in playing games of the end user. The resulting correlations with MOS show once again a heavy dependency on the type of game.

Wang et al. [23] propose a piece-wise linear metric for interactive GVS, based on bitrate, framerate, resolution, the video codec characteristics, available bandwidth, delay, jitter and packet loss. This metric is then polynomially fitted to subjective MOS, resulting in a 0.92 correlation. These results are based on a rather limited dataset, however, that consists of only three different games (a RPG, a

racing game & a sports game), 2 resolutions, 2 framerates and 2 delay values combined with a larger set of 8 Packet Loss Ratios (PLRs), ranging from 0 to 8 %.

Zadtootaghaj et al. [24] propose a metric for interactive GVS by including Peak Signal-to-Noise Ratio (PSNR) alongside bitrate and framerate in the set of input features. They propose a linear combination of polynomial and exponential expressions, resulting in correlations between 0.89 and 0.91 to MOS, depending on the game type. It should be remarked, however, that the study has only been performed on a dataset consisting of two games (*Grand Theft Auto V* & *Project Cars*), 2 bitrates and 4 framerates, thus only providing a total of 16 testing conditions. In addition, it has to be noted that the inclusion of PSNR in the feature set makes the model FR by construction, making it unusable in live-streaming scenarios.

On the topic of passive GVS, Barman et al. [25], perform a correlation analysis of multiple FR metrics (PSNR, Structural Similarity Index (SSIM), Visual Information Fidelity (VIF) and Video Multimethod Assessment Fusion (VMAF)) to MOS, for videos with different degrees of compression. Their results show average correlations of 0.70, 0.52, 0.67 and 0.88 respectively. In a second study [26], they present a modelling approach in which a Support Vector Regression (SVR) is used to forge a quality metric based on a set of NR-metrics. This set includes features such as blockiness and noise combined with a spatio-temporal characterization of the particular game. The model is trained on sequences annotated with VMAF, which is a FR-metric created by *Netflix* [29], resulting in a 0.98 correlation. Furthermore, a correlation of 0.89 with MOS is obtained. It should be noted, however, that the VMAF to MOS correlation tends to differ from a high 0.97 correlation value to a more limited 0.8 correlation, depending on the encoding characteristics. The encoding bitrate is a factor that has not been researched as the main focus lays on the resolution. Therefore, one might question how representative the end-user quality assessment of the first study [25] is towards scenarios with heavy compression. Furthermore, it highlights the need for a more stable FR-metric tailored to the specific context of GVS. In a third, subsequent study [27], they present two quality predicting models for passive GVS, *i.e.* No-Reference Gaming Video Streaming Quality Index (NR-GVSQI) and No-Reference Gaming Video Streaming Quality Estimator (NR-GVSQE). The NR-GVSQI model is based on an ANN and is used to predict subjective MOS scores. It takes a set of 15 NR features as an input, covering bitrate, resolution, Spatial Information (SI), Temporal Information (TI) and a variety of both spatial and temporal distortion metrics. Their model is evaluated on two separate datasets and shows correlations of 0.87 and 0.89 to MOS. The NR-GVSQE model uses a SVR to predict objective FR VMAF scores. A similar set of features is taken as the input. The results show correlations up to 0.97 to VMAF and 0.91 to MOS.

Göring et al. [28], at last, present a NR approach for passive GVS, based on a set of features including the Fast Fourier Transform (FFT), SI, TI, blockiness, blockmotion, staticness, temporal features based on cuboid slices, Natural Image Quality Evaluator (NIQE) and Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE). These features are fed to a temporal pooling entity based on mean and standard deviation calculations. The resulting values are used as input to a Random Forest (RF) benchmarked with VMAF to build the actual model. Their results show correlation values up to 0.96 to VMAF and 0.91 to subjective MOS as evaluated on the publicly available dataset of Barman et al. [19].

As can be concluded from the above overview, the amount of related work within the topic of GVS quality assessment is rather limited. In addition, the majority of the presented studies tends to research end-user perception in interactive rather than passive GVS. Visual satisfaction is an important part of the user's QoE in interactive scenarios, along with delay. Therefore, it gives important indications for the visual perception of its passive counterpart. However, due to the fundamental differences between both scenarios, both perceptions cannot just be equated. As the user acts as a spectator rather than a player in passive GVS, it can be assumed that he/she keeps a more general overview of the received content in comparison with the active user. The latter probably focuses more on certain aspects within the content of the stream, such as his/her personal avatar. Therefore, it is useful to investigate how this perception changes for the passive case of GVS.

A second, important conclusion is that no objective, FR-metric seems to exist that is known to provide satisfying results towards the assessment of end-user perception of GVS. Instead, objective metrics constructed for the natural video case are often applied for game-related, synthetic content despite their varying performance for this particular case.

Furthermore, it has to be pointed out that a significant part of the studies that try to relate the computationally less expensive NR/RR-metrics to an objective or subjective benchmark, are almost all focusing on bitstream-based methods. In addition, these studies show either limited performance or high performance on limited datasets. Furthermore, the role of the human visual system (HVS) characteristics within the end-user perception is rarely investigated.

At last, it has to be emphasized that most studies claim that the type of game under scrutiny plays an important role in the end-user perception of the content as well as the objective modelling of these subjective scores. In what way this game-type should be defined and incorporated within the modelling approach is an open research question, however.

Therefore, this work researches the establishment of an

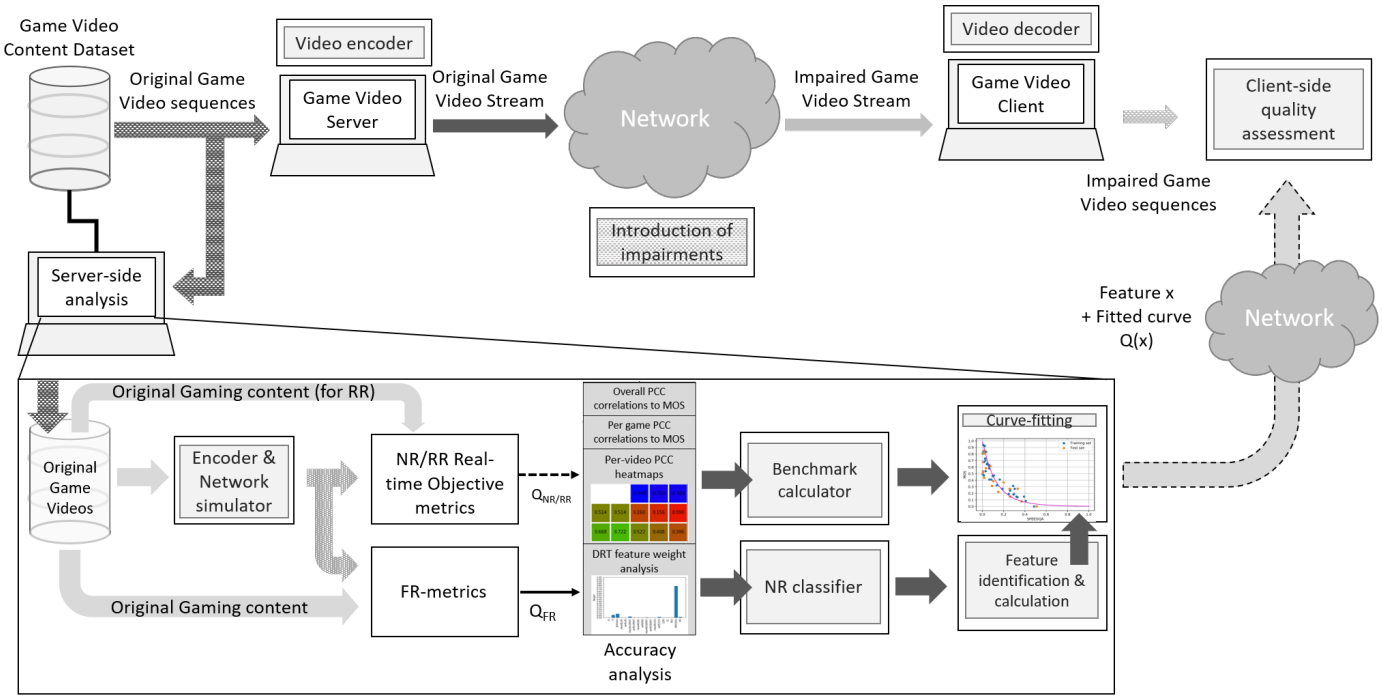


Figure 1: Schematic illustration of the presented end-to-end solution.

end-to-end solution for the real-time quality assessment of streamed game videos. To this end, a low-complexity psychometric curve-fitting approach (incorporating the characteristics of the HVS) is proposed that is based on a set of NR and RR features. These features include both pixel and bitstream-based NR and RR metrics. The best suited metric for the fit is selected based on a low-complexity game classifier. Afterwards, the obtained model's performance is benchmarked against a customized, FR GVS-metric.

3. A psychometric curve-fitting approach for quality assessment of passive GVS

Figure 1 presents an overview of the theoretical end-to-end solution. As can be seen, the live game video sequences, which are also recorded and stored in a database on the server side, are encoded to a game video stream which is sent by the *Game Video Server* over the network to the *Game Video Client*. As a result of this transmission, visual impairments occur that could negatively influence the end-user's perceived quality. In order to model this quality degradation, an offline *Server-side analysis* of the stored game video content is performed to establish an appropriate curve-fitting model. To this end, the impairments resulting from a multitude of encoding and network circumstances are simulated by means of an *Encoder & Network simulator*. Upon the resulting, impaired video sequences, a multitude of computationally efficient real-time NR/RR features is calculated alongside a set of more complex FR metrics. Both are subjected to a performance evaluation, including analysis towards correlation and DRT feature weights. Based on the results of this analysis, a game classification method can be derived using a limited amount of NR-features. Afterwards, the most suiting NR/RR feature for curve-fitting is selected based on the resulting class. A psychometric curve can than be fitted through this particular feature against a quality benchmark chosen based upon the performance of the FR-metrics. The particular feature x along with the resulting, fitted curve $Q(x)$ is then sent to the client-side where it can be used for real-time quality assessment. This is done by recalculating x on a regular base upon the received live-stream after which $Q(x)$ can be evaluated in the resulting value to obtain an estimation of the client-side perceived quality.

The remainder of this Section provides an in-depth description of the working principles of both the online client-based quality assessment method and the offline server-side pre-processing methodology.

3.1. Real-time model assessment

Figure 2 shows the typical, general relationship between a certain impairment and the end-user's perception. Three major areas can be distinguished [30]:

- *Area 1 (Constant perception):* The occurring degree of impairment is low enough, such that a slight increase or decrease of this particular impairment has little to no effect on the perceived quality [30].
- *Area 2 (Decreasing perception):* The degree of impairment grows above a certain threshold x_1 after which the perceived end-user quality tends to decrease rather fast [30].

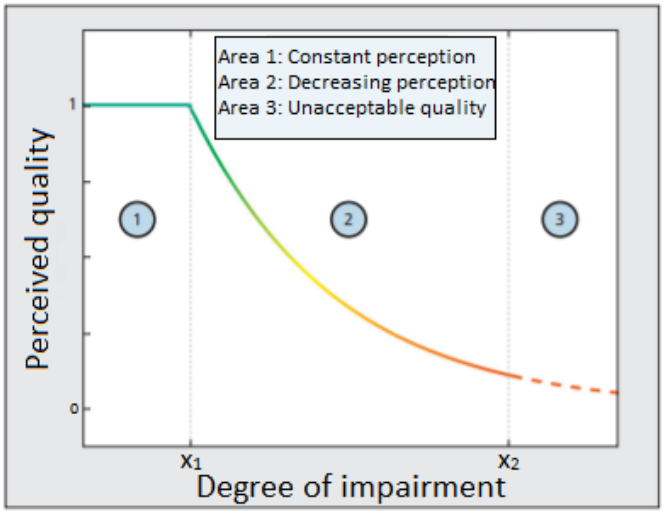


Figure 2: Typical, general relationship between the degree of impairment and the perceived quality in a multimedia service[30].

- *Area 3 (Unacceptable quality):* : After the degree of impairment exceeds a certain threshold x_2, the quality experienced by the user becomes so low that the user considers it as being unacceptable and decides to leave the system [30].

Due to this typical, psychometric impairment-quality relationship, sigmoidal curves are often applied to these kinds of curve-fitting approaches. This is because of their asymptotic behaviour in the neighbourhood of either very high or very low values of the variable under scrutiny, and their fast decreasing/increasing characteristic for medium values of the selected feature. Alternatives include exponential functions, whereas the values are clipped at the minimum and maximum quality value that can be obtained, or even a straightforward linear fit in some cases.

Equation 1 shows Equations for the sigmoidal ($Q_s(x)$), exponential ($Q_e(x)$) and linear ($Q_l(x)$) functions.

$$\begin{aligned} Q_s(x) &= \frac{1}{(1+\alpha_s \exp(-\beta_s x))^{\frac{1}{\gamma_s}}} \\ Q_e(x) &= \alpha_e \exp(\beta_e x) \\ Q_l(x) &= \alpha_l x + \beta_l \end{aligned} \tag{1}$$

Hereby, the linear and exponential curve are clipped at 0 and 1. Note that the exponent $\frac{1}{\gamma_s}$ in the denominator of the sigmoid curve has been added to allow to tune the speed at which the curve is increasing from 0 to 1. The weights α_i, β_i and γ_i, $i = l, e, s$, are calculated during the curve-fitting procedure such that the difference between the curve and the training points, *e.g.* expressed in Mean Squared Error (MSE), is minimized. It can be observed that these curves provide output values between 0 and 1 by definition, according to the normalized benchmarks. As a result, no constant terms are included in the Equations, as all other properties of the curves can be tweaked using the already provided parameters.

Heuristics are needed, however, to apply a curve-fitting approach in practice. The placement of the curve within the impairment-perception plane, for instance, is often depending on the dataset, the game type or even the particular sequence as human perception tends to be relative to a certain reference, rather than an absolute value. Therefore, in order to effectively assess the quality of a live game stream on the client-side, a starting point, *i.e.* an *anchor point*, of the psychometric curve is needed. To this extent, the following approach is proposed, which is illustrated in Figure 3. If one assumes that the game sequence labeled as *Sequence 2* is the current, real-time behaviour of the live-stream, while *Sequence 1* is a recording of the same game (but a different part of the content) saved within the server's database, the former's psychometric curve can be estimated on the latter's one. By measuring one or two seconds of the particular feature on the live-stream and comparing the feature value with a similar sequence of the particular game in the server's database, an estimation can be made of the shift that should be applied to the fitted curve (through the server data) to obtain the estimated curve of the real-time stream.

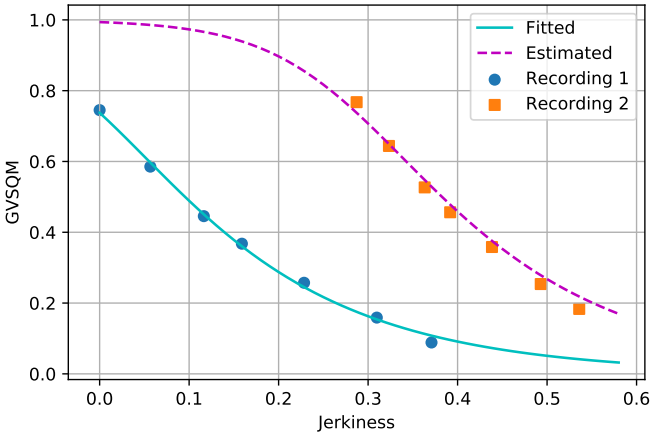


Figure 3: Illustration of the curve shift estimation procedure.

Another phenomenon that can occur is a dependency of the curve on the resolution. This can be easily solved, however, by fitting and sending one curve per supported resolution. On the client-side, the appropriate curve can easily be selected based on the resolution of the received stream. Specific experimentation and results concerning the curve-fitting procedure are provided in Section 6.

3.2. Server features accuracy and psychometric classification

Given the presented curve-fitting approach of the previous Section, the question remains how to select a certain feature for a particular game to fit a curve through. Therefore, a server-side pre-processing step is taken. On the server-side of the presented architecture, low-complexity features are calculated on a regular basis upon the frequently arriving game recordings within the database. These can be bitstream-based methods such as bitrate and

resolution as outputted by the encoder and/or network simulator, but also pixel-based features calculated upon the distorted content are possible. Furthermore, a set of FR-benchmarks are calculated in order to be able to train the model.

A thorough performance analysis of these NR/RR features and FR metrics in terms of correlation with subjective scores (*e.g.* MOS) is performed. These correlations are calculated for each metric on the dataset as a whole, as well as per game and for the multiple bitrate-resolution encoding pairs to obtain a genuine view on the performance of these metrics over the different conditions. This is also schematically shown in Figure 1. The resolution-bitrate performance is visualized in *correlation heatmaps*. These heatmaps are created in a cumulative fashion. This means that the correlation value shown at resolution r and bitrate b is calculated over all data points d for which resolution $r_d \leq r$ and bitrate $b_d \leq b$. This is done to make sure that enough data can be found to calculate a reliable correlation value upon, as correlations calculated on too little data might draw a distorted picture of the underlying relationships. In addition, it allows to analyse existing trends which could help to provide additional understanding on how far they correlate when evaluated on bitrates and resolutions. One heatmap per video is created to be able to assess the per content correlation over these parameters. The correlation values themselves are evaluated using the Pearson Correlation Coefficient (PCC), which measures the linear correlation between two sets of N data points $\vec{\mathbf{x}}$ and $\vec{\mathbf{y}}$ as the ratio between the sample co-variance of the two variables to the product of their sample standard deviations [31], *i.e.*

$$\mathrm{PCC}(\vec{\mathbf{x}}, \vec{\mathbf{y}}) = \frac{\sum_{i=0}^{N-1}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=0}^{N-1}(x_i - \bar{x})^2}\sqrt{\sum_{i=0}^{N-1}(y_i - \bar{y})^2}} \quad (2)$$

with x_i and y_i the individual data points of the two sets $\vec{\mathbf{x}}$ and $\vec{\mathbf{y}}$, respectively, and $\bar{x}$ and $\bar{y}$ their respective means. The resulting value is a decimal number between -1 and 1, where -1 and 1 indicate a strong negative or positive correlation respectively, and 0 no correlation at all [31]. Note that the Spearman's Rank Order Correlation Coefficient (SROCC) provides a valuable alternative to PCC. The latter is preferred, however, as the presented analysis aims at providing insight in the linear trends between variables rather than the monotony of their relationship.

In order to obtain additional insight in which NR/RR features have the highest impact to distinguish between high and low end-user quality, DRTs are trained upon the recordings of each game. This is also illustrated in Figure 1. Hereby, the subjective MOS scores are used as the benchmark. The DRT searches for the input feature and the splitting point within this feature's range that minimizes a certain error function over the resulting subsets of the data. The chosen error function $f(S_1, S_2)$ is the sum of the MSEs between predictions and benchmarks of both subsets S_1 and S_2, *i.e.*

$$f(S_1, S_2) = \frac{1}{\#S_1}\sum_{y_i \in S_1}(y_i - \bar{y_1})^2 + \frac{1}{\#S_2}\sum_{y_i \in S_2}(y_i - \bar{y_2})^2 \quad (3)$$

with $\#S_1$ and $\#S_2$ the resulting number of data points in subsets S_1 and S_2, respectively, and $\bar{y_k}$ the average of the benchmarks in S_k, $k = 1, 2$, functioning as an output for that particular subset. The same approach can recursively be followed until only one data point remains within each subset. These are called the *leafs* of the tree. When used as an actual prediction model, this behaviour is unwanted as it results in heavy overfitting behaviour and therefore in poor performance on unseen data. Therefore, a lower bound is often set to the degree of *impurity*, *i.e.* the error function, to deny a subset of being split if this would mean that the set bound would be surpassed. As the DRTs are, in first instance, only used to gain insight in the data rather than an actual modelling approach, however, trees can be allowed to grow completely.

Based on the results of this analysis, classes of games can be identified with similar relevant features when it comes to distinguishing between higher and lower benchmarks using DRTs. The games within each class can then be compared in terms of their spatio-temporal characteristics in order to obtain an objective classifier using only a few, low-complexity NR-features. Low-complexity, unsupervised clustering algorithms, *e.g. k-means*, can typically be used to this extent. More detailed results of this approach are provided in Section 5.2.

To this extent, a need arises for an objective, FR metric that could provide overall good performance. Therefore, based on the results of the performance analysis of standard, natural-video FR metrics for the particular case of GVS, the decision can be made whether one or more existing metrics prove to be accurate enough for application in the GVS case. If not, a new metric could be forged out of the existing metrics, whereas the new metric should show good performance both in terms of correlation accuracy and stability over the multiple distortion conditions. The resulting FR-metric can then be used to benchmark a dataset in order to evaluate the performance of the proposed model. In the most general case, this customized metric should take the psychometric characteristics of human perception, as discussed in the previous Section, into account, i.e.

$$\mathrm{FR}_{\mathrm{GVS}} = \sum_{i=0}^{N-1} \alpha_i \cdot f_i(\mathrm{FR}_i) \quad (4)$$

with $\{\mathrm{FR}_i | i = 0, ..., N-1\}$ a set of existing FR quality metrics and $\{f_i | i = 0, ..., N-1\}$ a set of psychometric functions. The latter can be either exponentials, logistic curves, combinations of both or even nothing more that the identity function. Further experimentation and results towards the analysis of existing metrics, as well as the creation of a customized metric are provided in Section 5.3.

Table 2: Overview of the Gaming VideoSET characteristics. The games and bitrates indicated in bold are provided with MOS [19].

Games	**Counter Strike: Global Offensive (CSGO)**, Diablo III (Diablo), Defense of the Ancients 2 (DotA2), **FIFA 17 (FIFA)**, **H1Z1: Just Survive (H1Z1)**, **Hearthstone (HS)**, Heroes Of The Storm (HOTS), **League of Legends (LoL)**, **Project Cars (PC)**, PlayerUnknown's BattleGround (PUBG), StarCraft2 (SC2), World of Warcraft (WoW)
Duration	30 s
Frame	30 fps
Encoding	CBR
Metrics	SpEED-QA, PSNR, SSIM, VMAF
MOS	90 videos, 5-point ACR scale, single-stimulus
Resolution [Pixels] & Bitrate [kbps]	• 640x480: **300**, 400, **600**, 900, **1200**, **2000**, **4000** • 1280x720: **500**, **600**, 750, 900, **1200**, 1600, **2000**, 2500, **4000** • 1920x1080: **600**, **750**, 1000, **1200**, 1500, **2000**, 3000, **4000**

Table 3: Set of calculated NR/RR features.

Type	Name	Acronym
NR-B	Encoding bitrate	BR
	Encoding resolution	RES
	Scene Complexity [32]	SC
	Level of Motion [32]	LOM
NR-P	Mean amount of blurriness [33]	MBLU
	Variance of blurriness [33]	VBLU
	Mean blur ratio [33]	MBLR
	Variance of the blur ratio [33]	VBLR
	Mean amount of noise [33]	MNO
	Variance of noise [33]	VNO
	Mean noise ratio [33]	MNOR
	Variance of the noise ratio [33]	VNOR
	Mean blockiness [34]	MBLK
	Variance of blockiness [34]	VBLK
	Spatial Information [35]	SI
	Temporal Information [35]	TI
	Jerkiness [36]	JER
	Variance of the Motion Intensity [36]	VMI
RR	SpEED-QA [9]	SPEEDQA

Table 4: Set of FR metrics being evaluated.

Name	Acronym
Peak Signal-to-Noise Ratio	PSNR
Structural Similarity Index	SSIM
Video Quality Metric	VQM
Video Multimethod Assessment Fusion	VMAF

4. Dataset

The dataset used for the research is the *Gaming VideoSET* [19]. It consists of 24 unimpaired recordings from 12 games (2 recordings per game), spanning different game genres [19]. These games are the following:

- *Counter Strike: Global Offensive (CSGO)*: Realistic FPS from 2012 in which two teams (terrorists & counter-terrorists) with conflicting goals, such as detonating/disarming a bomb or capturing/protecting a flag, fight each other [19].
- *Diablo III (Diablo)*: A 2012 fantasy RPG in an isometric perspective in which players fight monsters in dungeons allowing them to upgrade their chosen character over time [19].
- *Defense of the Ancients 2 (DotA2)*: An online multiplayer game from 2013 in an isometric perspective, taking place in a fantasy battle arena and very popular in eSports competitions. Each player controls a single avatar from a five-a-side team. Two teams fight one against another in order to destroy the opponents base as fast as possible [19].
- *FIFA 17 (FIFA)*: One of the famous realistic soccer-simulating games from Electronic Arts, released in 2016. Players can play soccer matches against the built-in Artificial Intelligence (AI) but also against other players, both offline and online [19].
- *H1Z1: Just Survive (H1Z1)*: Survival game from 2015, situated in a realistic, post-apocalyptic setting and third-person view. Goal is to take back control in a world being overtaken by zombies [19].
- *Hearthstone (HS)*: Spin-off tabletop card game from the popular fantasy World of Warcraft (WoW) games, released in 2014. This title features a turn-based card game in a fantasy setting [19].
- *Heroes Of The Storm (HOTS)*: Real-time action strategy game from 2015 in an isometric perspective, situated in a fantasy battle arena setting. Similar to DotA2, 2 five-a-side teams try to destroy each other's bases [19].
- *League of Legends (LoL)*: A fantasy battle-arena game in an isometric perspective, released in 2009, and similar to both DotA2 and HOTS. It is very popular in eSports and one of the most played games worldwide. Furthermore, it is intensively watched on passive GVS platforms such as Twitch [19].
- *Project Cars (PC)*: A very realistic, racing simulator game from 2015, offering a wide choice in possible cars, tracks and camera perspectives [19].
- *PlayerUnknown's BattleGround (PUBG)*: An online multiplayer battle game from 2017. The game is situated in a first-person realistic setting, with the concept consisting of a large group of players (max. 100) being dropped randomly on an island. Goal is to find weapons and equipment in order to eliminate other players. The last one standing wins the game [19].
- *StarCraft2 (SC2)*: Fantasy real-time strategy game in an omnipresent perspective, released in 2010 and popular in eSports competitions. Players create bases

and units and try to destroy the opponent's equipment [19].

- *WoW*: A well-known fantasy-style Massive Multiplayer Online Role-Playing Game (MMORPG) from 2004 in an isometric perspective. Each player creates his own avatar that can evolve over time by completing certain tasks in both competitive and co-operative settings [19].

For each game, both recordings span a 30-second timespan and were losslessly captured in 1080p resolution and 30fps in RGB-format. Afterwards, they were converted to YUV-format. Out of these 24 recordings, 576 impaired MP4-encoded game video streams were created using 24 different bitrate-resolution combinations [19], as shown in Table 2. The encoding is performed at Constant Bit Rate (CBR), which is common behaviour for GVS, because streamed gaming videos often have fast alternating periods of high action and rather static gameplay. Variable Bit Rate (VBR) encoding could therefore result in end-user stall of the stream, which is highly unwanted [37]. In addition, each of these impaired sequences was annotated with the frequently used FR-metrics PSNR, SSIM and VMAF and an additional RR-metric, being Spatial Efficient Entropic Differencing for Quality Assessment (SpEED-QA). Moreover, a subset of 90 video streams was annotated with subjective MOS (indicated in bold in Table 2) [19]. A total of 25 test subjects, with a median age of 29 and covering different demographic backgrounds, scored the sequences in a single-stimulus approach on a 5-point Absolute Category Rating (ACR) scale [19].

5. Experimental accuracy analysis of OVQMs

This Section investigates the accuracy of objective metrics for GVS, based on the methodology presented in Section 3.2. To this extent, the small dataset consisting of 90 entries annotated with the ground-truth MOS is used. First, an overview is given of the features and FR quality metrics being calculated (Section 5.1). Next, the obtained results for the NR/RR features are discussed after which a game classifier is proposed (Section 5.2). Furthermore, a similar accuracy analysis is performed and discussed for the FR metrics as well (Section 5.3). Based on these results, a custom, objective metric suited to GVS is constructed. In Section 5.4, at last, a brief summary of the most important conclusions of this analysis is given.

5.1. NR/RR

Features & FR quality metrics

Next to the encoding bitrate, resolution and the SpEED-QA RR-metric already included in the dataset, an additional set of features is calculated to represent each video. This resulting set of features is presented in Table 3. First, the game video streams are analyzed in terms of their spatio-temporal information. This is done both on the pixel-based level (NR-P) in terms of Spatial (SI) and Temporal Information (TI) and on the bitstream level (NR-B) based on Scene Complexity (SC) and Level of Motion (LoM) [32]. The last two can be directly obtained from the *FFmpeg*-client [38]. Furthermore, a set of frame-by-frame pixel-based NR-metrics (results of which are considered as features) is calculated upon the distorted streams. These metrics include measurements of motion, blurriness, noise, blockiness and jerkiness. Each of these features, as well as the pre-calculated FR quality metrics, are normalized to the $[0, 1]$ interval in order not to favour one metric over another in distance based methods.

In addition to PSNR, SSIM and VMAF, the Video Quality Metric (VQM) [39] is calculated over each video sequence using the open-source *MATLAB* implementation provided by the Institute for Telecommunication Sciences (ITS) [40]. As VQM is intrinsically a measurement of *quality degradation* between 0 and 1, each of the videos is annotated with $1-$VQM to obtain a metric of *quality*, consistent with the other FR metrics. For the remainder of this work, every reference to VQM, actually means $1 - \text{VQM}$. The resulting set of objective FR metrics is shown in Table 4.

5.2. Evaluation of NR/RR features

In Table 5 the most important results of the correlation analysis are given per game and overall for the considered NR/RR-features to MOS. First of all, it is worth mentioning that SpEED-QA shows the strongest overall correlation with a PCC of -0.761. It is interesting to see, however, that SpEED-QA shows reasonable results for all games except HS. Only a value of -0.300 is obtained for this particular game. On the other side, strong correlations can be seen for SI, SC and MNO (0.968, 0.942 and -0.959 respectively), while the performance of these features is far more limited for the other games. Opposite behaviour is observed for JER. To further analyze this behaviour, the bitrate-resolution heatmaps are interpreted as well. In Figure 4, a subset of the NR/RR features with strong correlation to either one or multiple games is shown. Bitrate-resolution pairs for which no correlation value is given indicate that the correlation was calculated over either 1 (undefined) or 2 (1 by definition) data points. As a result, bitrates 300 and 500 are not included in the Figure (both only have one data point per game). They are included in the calculation of the correlation values, though. Both JER and SpEED-QA, for example, show strong negative correlations over all conditions for all games except HS. For higher resolutions of HS, correlation is almost nonexisting. Remarkably enough, HS even shows positive correlation for the lowest 640×480 resolution. On the other hand, high and low overall correlations can be noticed for the SC and MNO features, respectively. These correlations are rather stable over the different encoding conditions, whereas PCC values of these features are heavily changing over the multiple bitrates and resolutions for the other games. These observations provide a strong indication that the type of game might be an important factor

Table 5: Overview of the PCCs, per game and calculated over the whole dataset, between the NR/RR-features and MOS. The best and worst performing one (in absolute value) per game and overall are indicated in blue italic and red bold, respectively.

Game	VBLU	MBLR	SI	TI	MNO	VNO	MNOR	SC	LOM	VBLK	JER	SPEEDQA
CSGO	-0.358	0.444	0.329	-0.352	-0.783	0.399	-0.589	0.398	0.491	**-0.267**	*-0.920*	-0.865
FIFA	*0.938*	0.774	0.669	-0.414	-0.352	**0.007**	-0.684	0.662	0.896	-0.796	-0.886	-0.841
H1Z1	**-0.065**	-0.436	0.094	0.434	-0.504	0.733	-0.699	0.435	0.556	-0.291	*-0.954*	-0.925
HS	0.092	0.931	*0.968*	-0.287	-0.959	0.923	**0.040**	0.942	0.660	-0.515	-0.092	-0.300
LoL	**0.077**	0.809	0.436	-0.548	-0.727	0.596	-0.565	0.631	0.675	-0.488	-0.725	*-0.814*
PC	-0.594	0.345	0.354	-0.412	-0.611	0.464	**0.192**	0.253	0.322	-0.701	*-0.941*	*-0.941*
Overall	-0.240	**0.060**	0.508	-0.291	-0.444	**0.060**	-0.308	0.572	0.530	-0.388	0.044	*-0.761*

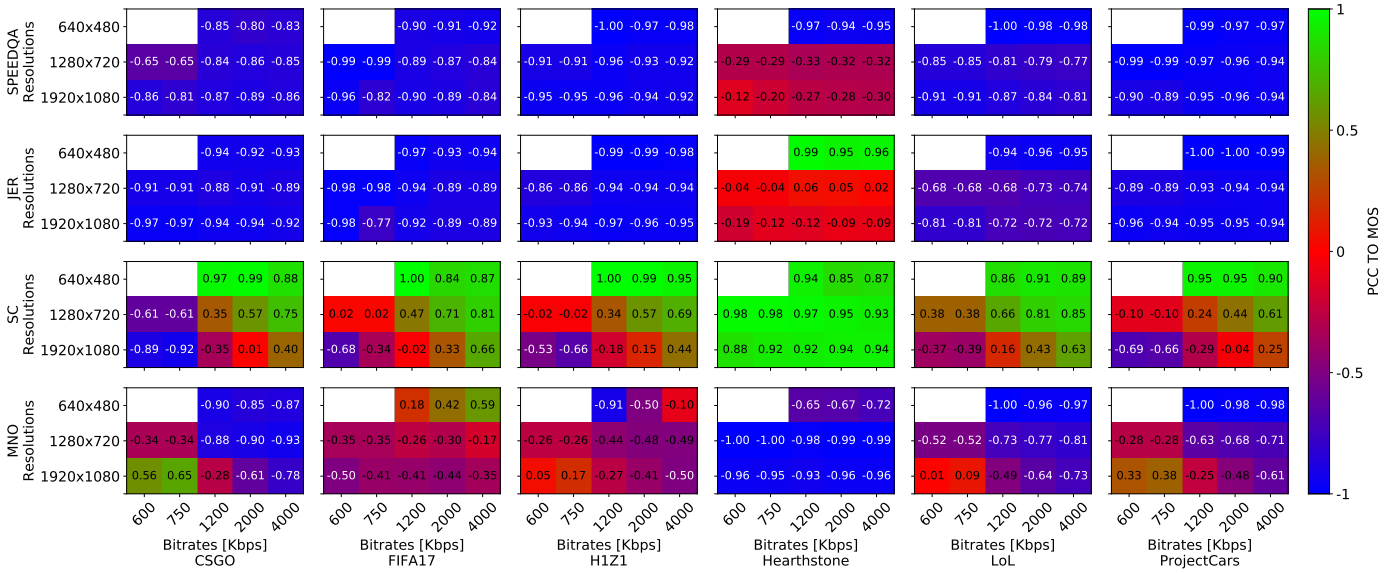


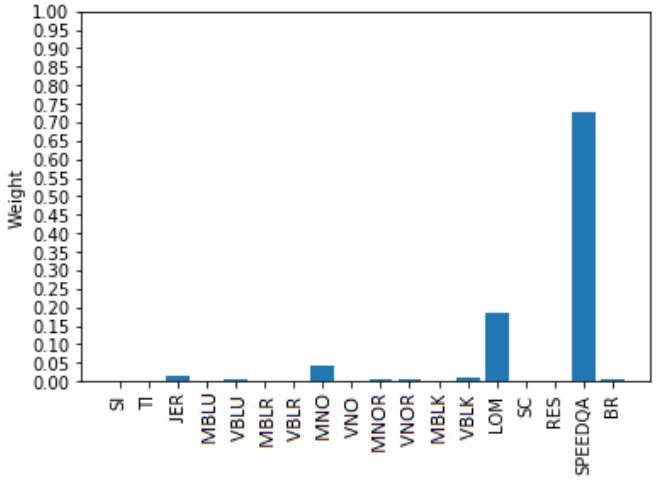
Figure 4: PCCs between four selected NR/RR features (SPEEDQA, JER, SC & MNO) and MOS for each of the six subjectively annotated games. Green means full correlation (PCC=1), dark blue full anti-correlation (PCC=-1) and red no correlation at all (PCC=0).

to be considered when modelling the end-user's perceived QoE. Moreover, it shows that, depending on the game, either temporal or spatial characteristics of the stream have a more important influence in the user's perception.

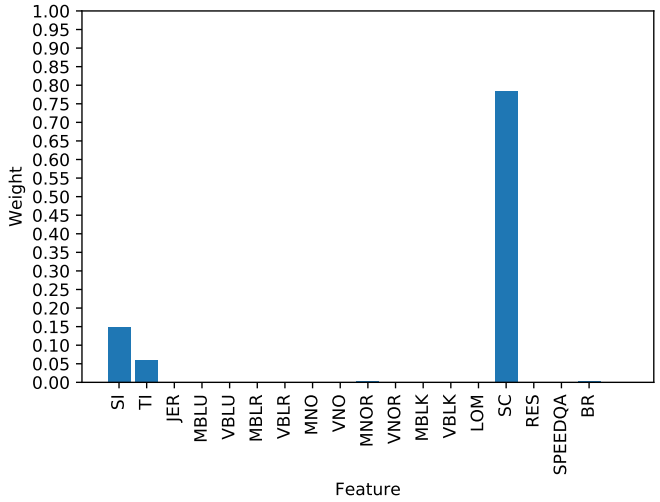
In addition to the above, a per-game DRT has been constructed, with MOS as a benchmark, and the weights put on the multiple features have been analyzed. The results show that each of the obtained trees tends to put a rather high weight (≥ 0.7) on one specific feature. This feature differs, however, depending on the game at hand. By means of illustration, the distribution of the feature weights for the particular cases of FIFA and HS is shown in Figure 5. For 4 out of 6 games, being FIFA, H1Z1, LoL and PC, SpEED-QA turns out to be the most important feature when differentiating between higher and lower subjective scores. For CSGO, JER seems to be the most important feature, while HS is heavily depending on SC. Note that the feature with the highest weight is not necessarily the feature with the highest PCC to MOS as indicated in Table 5. This is because the presented DRT uses MSE as the error function (Equation 3). As high PCC does not necessarily imply low MSE and vice versa, slight differences between both approaches might occur. Based on the obtained results, a heuristic distinction between three game classes can be made.

- Class 0: HS
- Class 1: FIFA, H1Z1, LoL & PC
- Class 2: CSGO

To obtain a more objective, game-independent classification, two spatial (SI & SC) and three temporal characteristics (TI, LoM and Mean Motion Intensity (MMI)) are investigated, where MMI is the mean-squared difference of adjacent frames averaged over the game video [36]. As can be seen from Figure 6, in which the data points are manually labeled with the classes derived from the DRT analysis, MMI on its own proves to be sufficient to provide a distinction between the multiple classes. A *k-means classifier*, with $k = 3$, that works solely on the MMI characteristic is therefore proposed as a classifying approach. The resulting classification, shown in Figure 7, indicates a 93,3 % resemblance with the manually annotated case. The fact that MMI proves to be sufficient to distinguish between different types of games, in combination with the leading features obtained from the DRT analysis leads to a rather interesting conclusion. The more motion is contained within a specific stream, e.g. CSGO, the more tem-



(a) FIFA (Class 1)



(b) HS (Class 0)

Figure 5: Illustration of the distribution of the DRT feature weights for the particular cases of FIFA and HS

poral artifacts such as jerkiness play a role in the end-user's QoE. Moreover, the (limited) difference between the manual annotation and the k-means classifier indicates that the encoding conditions itself also play their role in this. Spatial detail, on the other side, takes the upper hand in games with little motion, as is the case for HS. SpEED-QA holds the middle between both as it includes both spatial and temporal characteristics, making it the ideal metric for video sequences with medium MMI. This is an interesting observation, as similar approaches for natural videos show far more complex relationships between NR/RR features and end-user QoE, often resulting in black-box, ML-based modeling approaches.

Note that the application of k-means throws two additional questions when scaled to real-life applications. The first question to be answered is at which frequency the clustering mechanism should be re-trained in order to adapt the classifier to newly added and/or deleted game video sequences. This is of course highly dependent on the rate at which gaming videos are added/removed from the server database. As long as the number of videos used as *training set* for the classifier is high enough relative to the total amount of videos in the database, no re-iteration of the algorithm is needed (as the influence on the cluster centers is assumed to be limited) and the classifier can be used as such. Once this ratio exceeds a certain threshold, retraining is required. Further experimentation on larger datasets is needed, however, to derive an exact value for this particular threshold. Next, it should be noted that the use of k-means introduces a dynamic threshold in the classification procedure. In some cases (*e.g.* when no low MMI games such as HS are present) this could lead to misclassification. Given the application of the proposed framework (*i.e.* streaming of gaming videos rather than actual cloud gaming), however, it is not too far fetched to assume that the server database will cover a rather large and varied set of gaming video genres and thus, MMI values.

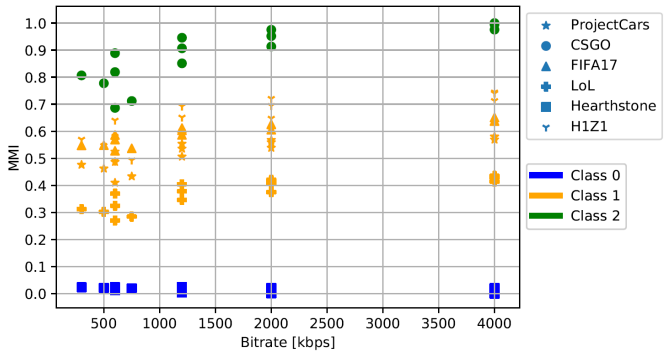


Figure 6: Two-dimensional distribution of the game videos within the bitrate-Mean Motion Intensity (MMI) plane. The data points are manually labeled with the classes derived from the DRT analysis. Blue is class 0, yellow class 1 and green class 2.

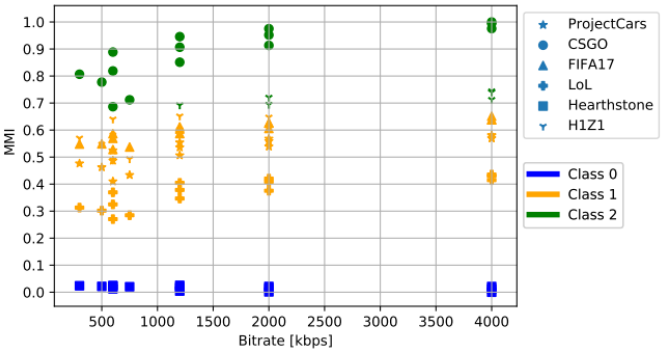


Figure 7: Two-dimensional distribution of the annotated classes of the game videos within the bitrate-MMI plane after k-means ($k = 3$) clustering. The resulting classification shows a 93,3 % resemblance with the manually annotated case.

5.3. *Evaluation of FR metrics*

A similar analysis is performed for each of the four FR-metrics, as shown in Table 6. It can be seen that each of them provides a reasonable overall correlation with

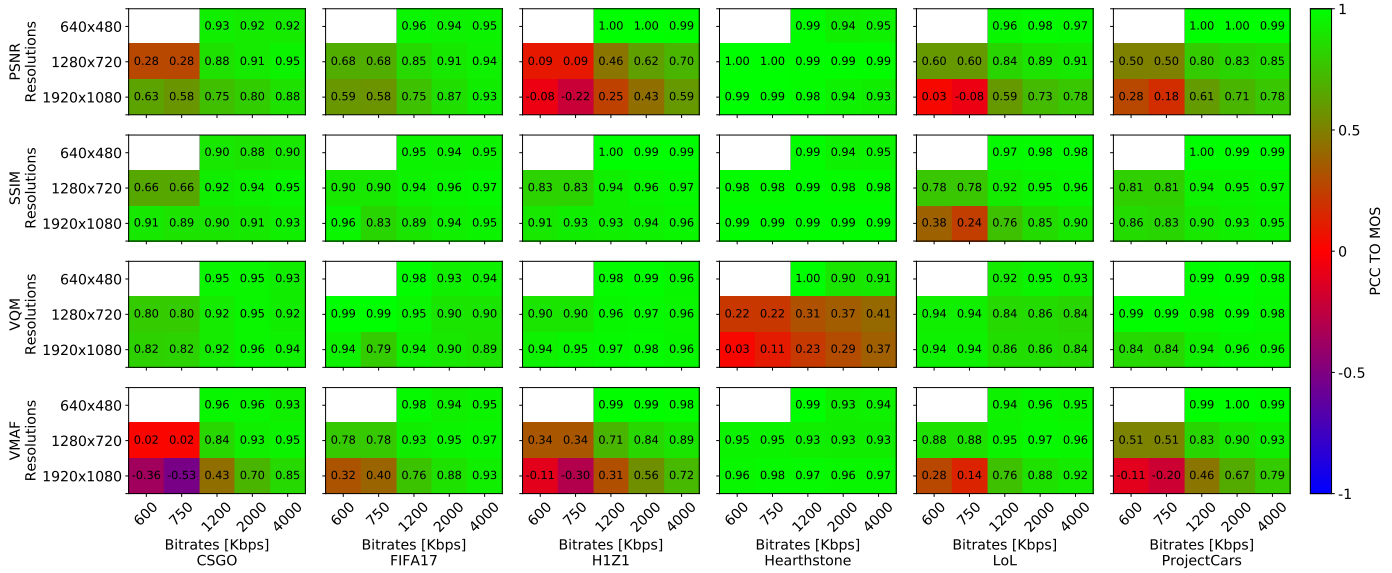


Figure 8: PCCs between the four considered FR metrics (PSNR, SSIM, VQM & VMAF) and MOS for each of the six subjectively annotated games. Green means full correlation (PCC=1), dark blue full anti-correlation (PCC=-1) and red no correlation at all (PCC=0).

Table 6: Overview of the PCCs, per game and calculated over the whole dataset, between the four FR-metrics and MOS. The best and worst performing metric per game and overall are indicated in blue italic and red bold, respectively.

Game	PSNR	SSIM	VQM	VMAF
CSGO	0.88	0.934	*0.94*	**0.847**
FIFA	0.926	*0.95*	**0.89**	0.934
H1Z1	**0.594**	0.961	*0.965*	0.765
HS	0.925	*0.987*	**0.368**	0.968
LoL	**0.781**	0.901	0.845	*0.919*
PC	**0.775**	0.953	*0.96*	0.79
Overall	**0.741**	0.79	0.825	*0.864*

MOS, with PCCs varying from 0.741 (PSNR) to 0.864 (VMAF). Furthermore, it is worth noting that for 3 out of 6 games, VQM tends to provide the best performance, hereby clearly outperforming VMAF for the game at hand. Deviant behaviour can again be observed for the HS game, with only a 0.368 correlation of VQM to MOS while its VMAF counterpart obtains a high 0.968 PCC. Somewhat surprisingly, SSIM also shows rather high, per game correlations. Sometimes, it is even performing better than both VQM and VMAF, as is the case for FIFA and H1Z1. This might be an indication of the fact that the structural information in a video sequence plays a far more important role towards end-user QoE for synthetic game content than is the case for natural videos.

Figure 8 shows the bitrate-resolution correlation results for each of the games. An interesting conclusion is that VMAF, the metric with the best overall performance, only tends to show this performance for high bitrates and low resolutions, while performance is dropping heavily for the opposite case. Again, HS shows to be the exception with

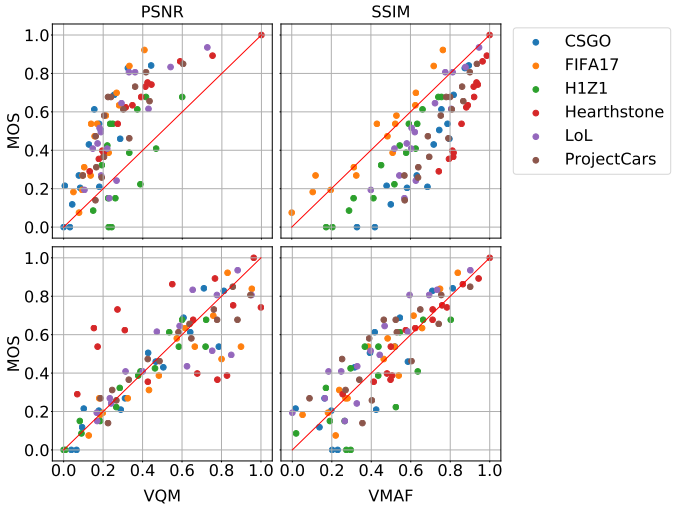
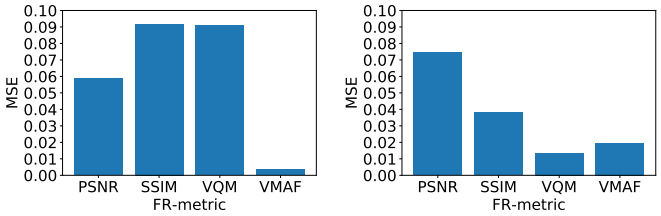


Figure 9: Scatterplots, showing the relationship between the four objective FR metrics and subjective MOS.

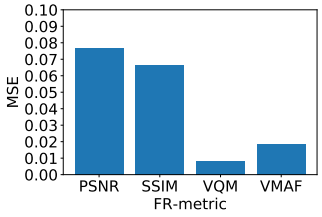
rather stable behaviour of the VMAF-metric. VQM, on the other hand, shows high and stable correlation values for all games except HS. For the latter, VQM shows a heavy drop in performance for higher resolutions. As could already be seen from Table 6, SSIM shows somewhat surprisingly to be the most stable metric over the encoding conditions, only showing a performance drop worth mentioning for the lowest bitrates and highest resolution of LoL.

PCC gives an objective indication on the linear relationship between two variables, but does not provide an indication on over- and underprediction, whatsoever. So, this behaviour is worth analyzing as well. Figure 9 shows scatterplots for each of the four considered FR metrics



(a) Class 0

(b) Class 1



(c) Class 2

Figure 10: Illustration of the per-class MSEs between the four objective FR metrics and subjective MOS.

towards MOS. It can be seen that PSNR tends to consistently underpredict the MOS scores, while SSIM shows the opposite behaviour with FIFA as its most important exception. VQM shows, apart from HS, rather accurate predictions for the lower regions of the MOS scores, but tends to drop in performance for higher ones. VMAF has the most accurate performance, although it can be noted that higher MOS-scores seem easier to predict than lower ones. Given the fluctuating behaviour of these metrics, both in terms of PCC and over/underprediction, it is of great interest to construct an accurate, objective score to be used as a benchmark such that it shows both accurate prediction results without over- or underpredicting while maintaining stability over a multitude of encoding conditions.

Therefore, the accuracy of the FR-metrics in terms of the MSE towards MOS is investigated per class as obtained using the MMI-classifier. The results, shown in Figure 10, illustrate that VMAF has a significantly lower MSE for the game within class 0, *i.e.* HS. For class 1 and 2, both VQM and VMAF show reasonable performance, Based on this behaviour, an objective metric, which is called the GVSQM, is proposed that equals a per-class weighted combination of VQM and VMAF. The following combinations turn out to minimize the MSE with MOS.

$$\text{GVSQM} = \begin{cases} \text{VMAF} & \text{if class} = 0 \\ 0.584 \cdot \text{VQM} + 0.416 \cdot \text{VMAF} & \text{if class} = 1 \\ 0.706 \cdot \text{VQM} + 0.294 \cdot \text{VMAF} & \text{if class} = 2 \end{cases} \quad (5)$$

As the calculation of the proposed metric relies on a prior classification of the game at hand, however, it cannot be used independent of the dataset being studied, which is unpractical. One can notice from Equation 5, though, that the relative weight of the VQM-metric within the GVSQM calculation tends to increase with increasing class number

Table 7: Overview of the per-class and overall performance of the GVSQM metric, both in terms of PCC and MSE to MOS.

Class	PCC	MSE
0	0.964	0.004
1	0.929	0.009
2	0.933	0.008
Overall	0.939	0.008

and thus with increasing MMI. Based on this observation, the GVSQM metric is made dataset-independent by using the MMI feature of the sequence at hand as a weight in the equation, *i.e.*

$$\text{GVSQM} = \text{MMI} \cdot \text{VQM} + (1 - \text{MMI}) \cdot \text{VMAF} \quad (6)$$

This metric shows an overall PCC of 0.939 and 0.008 MSE to MOS. The performance metrics separated per class are similar, as can be seen from Table 7.

Furthermore, as can be seen from Figure 11, GVSQM shows much more stable behaviour over the multiple games, bitrates and resolutions as was the case for PSNR, VQM and VMAF (Figure 8). Only for SSIM, similar stability can be noticed. However, SSIM is heavily suffering from overprediction, as is illustrated in Figure 9. This is not the case for GVSQM, though, as one can notice in Figure 12. In addition, GVSQM shows similar behaviour for both the lower and higher regions of the subjective MOS, contrary to VQM and VMAF. As such, GVSQM proves to be a more accurate and stable metric for GVS in comparison with existing state-of-the-art video metrics. This allows GVSQM to be used as a benchmark in real-world GVS quality assessing frameworks, as the real-time collection of MOS is infeasible.

5.4. Conclusion

To summarize, it could be stated that a straightforward psychometric curve-fitting approach through one specific feature should be sufficient to model the ground-truth quality. This feature is either SC, SpEED-QA or JER, for classes 0, 1 and 2 respectively, given the high weights they receive from the DRT analysis. The class of the game video recording at hand can be determined based on the MMI based classifier, derived in Section 5.2.

6. Evaluation of the NR/RR psychometric curve-fitting approach

This Section provides the evaluation of the proposed psychometric curve-fitting approach. This is done in four, subsequent phases. First, accuracy is evaluated by a psychometric fit of the selected features (*i.e.* SC, SpEED-QA and JER) directly to MOS (Section 6.1). Second, this subjective benchmarking is evaluated against objective, FR metrics (Section 6.2). Both evaluations are performed on the small, subjectively annotated dataset of 90 records. Finally, the scalability of the approach is evaluated upon the

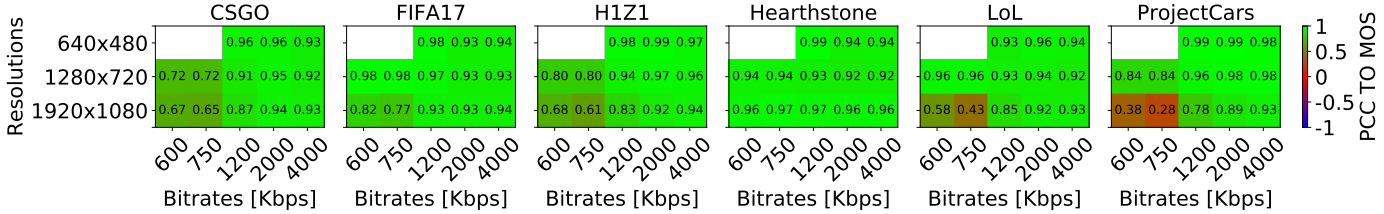


Figure 11: PCCs between the GVSQM metric proposed in Equation 6 and MOS for each of the six subjectively annotated games. Green means full correlation (PCC=1), dark blue full anti-correlation (PCC=-1) and red no correlation at all (PCC=0).

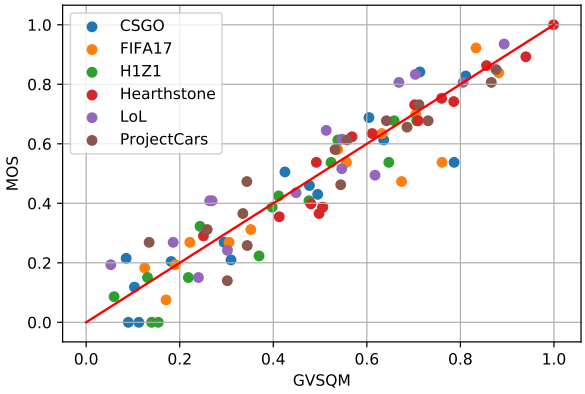


Figure 12: Scatterplot showing the relationship between the objective, GVSQM metric and subjective MOS.

full dataset of 576 datapoints, using an objective benchmark (Section 6.3). At last, a comparison of the presented curve-fitting approach with two other prediction mechanisms is provided (Section 6.4).

For the first three phases, both PCC and MSE are included to assess the performance of the proposed model. This is done because PCC gives information on the linear relationship between two variables, while MSE is a measurement for the amount of over- or underprediction when combined with PCC. These metrics are gathered using *stratified k-fold Cross-Validation (CV)*. Unlike standard k-fold, the data points within each fold are chosen such that they are evenly distributed over the 0 to 1 scale of the (normalized) benchmark. This is done to avoid that the particular model is fitted on data points falling within a small interval of the benchmark range, therefore resulting in low performance on other benchmark intervals [31]. To ensure this stratified behaviour, the benchmarks $l_i \in [0, 1]$ are mapped to n bins $b_i = [\frac{i}{n}, \frac{i+1}{n}], i = 0, ..., n-1$, with $n = 4$ in this particular case. Afterwards, the data points within each fold are chosen at random, but such that each bin is more or less evenly represented.

6.1. Accuracy analysis: benchmarking against MOS

To gather insight in the accuracy and applicability of the proposed curve-fitting procedure, the approach is first benchmarked directly to the ground-truth quality, *i.e.* MOS. In Table 8 the resulting performance for each of the classes and each of the investigated psychometric curves

Table 8: Overview of the obtained PCC and MSE scores towards MOS by applying a psychometric curve fitting approach

	Linear		**Exponential**		**Sigmoid**	
Class	**PCC**	**MSE**	**PCC**	**MSE**	**PCC**	**MSE**
0	0.955	0.007	0.929	0.026	0.963	0.004
1	0.882	0.016	0.904	0.014	0.904	0.014
2	0.978	0.021	0.964	0.026	0.982	0.019
Overall	0.873	0.016	0.852	0.019	0.883	0.013

(linear, exponential and sigmoidal) is shown, both in terms of PCC and MSE as averaged over the multiple test folds. It can be seen that a sigmoid function is showing the best performance for each of the classes, although leveled by the exponential curve for class 1. Especially for class 0, a curve fitting approach on a single feature (SC in this case) shows rather high performance with an obtained MSE as low as 0.004. In Figure 13, a scatterplot showing the relationship between the predicted and true MOS using the best performing, sigmoidal curve fitting approach is shown. To this extent, the prediction values for each test fold and each class are collected and displayed using different colors per class and different symbols per game.

6.2. Robustness analysis: benchmarking against objective metrics

As subjective scores are typically not freely available for a large set of video recordings, the curve-fitting is typically performed against an objective benchmark. Therefore, the outcome of this approach in relation to the ground-truth MOS is analyzed in this Section.

In Table 9, the obtained PCC and MSE to MOS are shown after applying a curve-fitting approach to each of the objective, FR benchmarks. This is done for each class, as well as overall. The curve is chosen to be a sigmoid, as it shows the best performance based on the results from the previous Section. The best and worst values for each metric and each class, as well as overall, are indicated in italic blue and bold red, respectively. It can be seen that both PSNR and SSIM show rather low performance in terms of MSE to MOS, with high values up to 0.100 in the worst case. This is a clear indication for over- or underprediction, which is in line with the results from Section 6.1. This can also be noticed from the scatterplots in Figure 14, showing the predicted, objective metrics against the actual

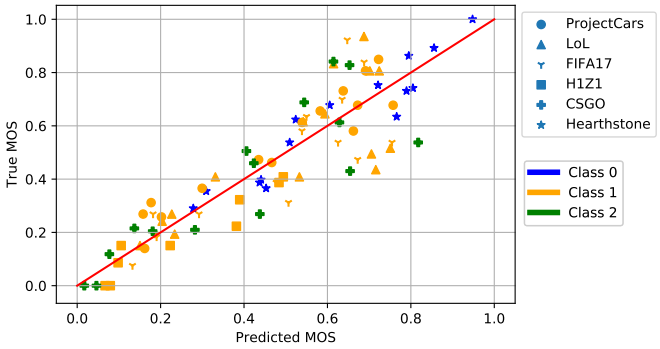


Figure 13: Scatterplot showing the relationship between the predicted and true MOS using a sigmoid curve fitting approach. The prediction values for each test fold and each class are collected and displayed using different colors per class and different symbols per game.

Table 9: Overview of the obtained PCC and MSE towards MOS by applying the psychometric curve-fitting approach to the objective benchmarks being indicated. The best and worst value of each metric are indicated per class, as well as overall, in blue italic and red bold respectively.

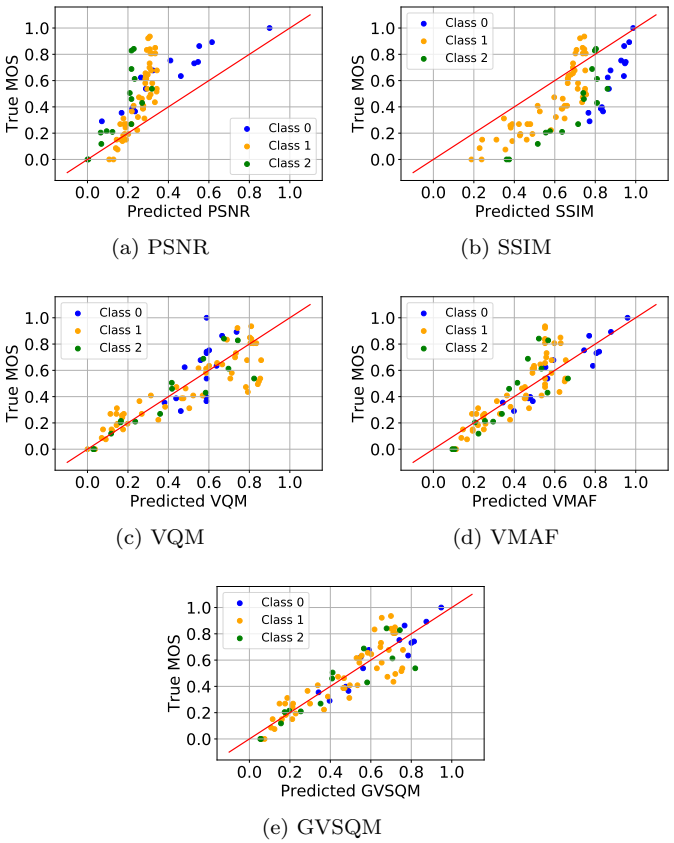
	Class 0		Class 1		Class 2		Overall	
	PCC	MSE	PCC	MSE	PCC	MSE	PCC	MSE
PSNR	0.928	0.059	0.860	**0.077**	**0.800**	0.092	**0.747**	**0.076**
SSIM	0.934	**0.096**	**0.853**	0.036	0.873	**0.100**	0.785	0.058
VQM	**0.675**	0.030	*0.887*	0.018	*0.912*	*0.012*	0.858	0.019
VMAF	*0.936*	*0.007*	0.866	0.019	0.870	0.021	0.871	0.017
GVSQM	*0.936*	*0.007*	0.884	*0.014*	0.909	0.013	*0.903*	*0.012*

MOS for each of the sequences. For both PSNR and SSIM, the same over-/underpredicting behaviour as indicated in Section 5.3 can be noticed.

When using VQM as a benchmark to predict MOS, on the other hand, acceptable performance is obtained. VQM is even showing the best results for class 2 both in terms of PCC and MSE, although closely chased by GVSQM. It has to be noted, however, that VQM shows a severe performance drop for class 0 games, which is in line with the results from the correlation analysis presented in Section 5.2. VMAF shows the most stable performance of the 4 natural video FR benchmarks, with acceptable PCC and MSE for each of the classes, as well as overall. GVSQM, however, at least matches this performance for every case and even shows improved overall performance with a 0.032 gain in PCC and 0.005 drop of MSE in comparison with VMAF. As a result, GVSQM shows to be the best benchmark, both in terms of linearity and accuracy of the prediction. This can also be concluded from the scatterplots in Figure 14.

6.3. Scalability analysis

As was proven in Section 6.2, GVSQM shows to be the most promising metric in terms of PCC and MSE to MOS. This Section investigates whether the proposed curve-fitting approach can be scaled to the full dataset using GVSQM as the benchmark. First, some adaptations to the previously presented curve-fitting will need to be



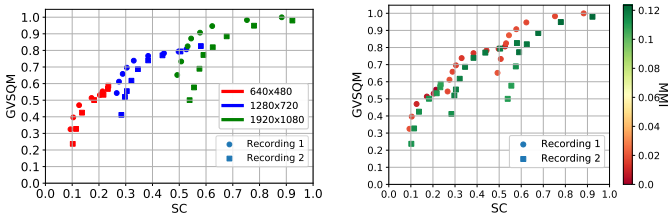
(a) PSNR (b) SSIM (c) VQM (d) VMAF (e) GVSQM

Figure 14: Scatterplots showing the relationship between the predicted, objective benchmark and the true MOS for that particular sequence, using a sigmoidal curve-fitting approach.

made after which the obtained results are discussed. Afterwards, DRTs and ANNs are applied to investigate how this approach performs in comparison with other common applied models in literature (Section 6.4).

In order to apply the curve-fitting approach, two additional problems are encountered on the full dataset that were previously hidden due to the small amount of data in the subjectively annotated set of 90 points. First of all, as can be seen from Figure 15a for the SC feature of the HS game, the curves show both a horizontal and a vertical shift depending on the resolution of the particular game stream, whereas higher resolutions induce higher GVSQM scores. Furthermore, a horizontal shift of the per-resolution curves can be observed that depends on the particular game video recording. More specifically, this shift is clearly a result of the difference in MMI between both recordings, as is shown in Figure 15b. Similar behaviour is concluded from other games and classes as well. Both problems are solved as described in Section 3.1.

Here, instead of MOS, we predict GVSQM scores applying the curve-fitting approach with the selected features (SC, SpEED-QA, and JER). Table 10 shows the resulting overall PCC and MSE of this curve-fitting approach,

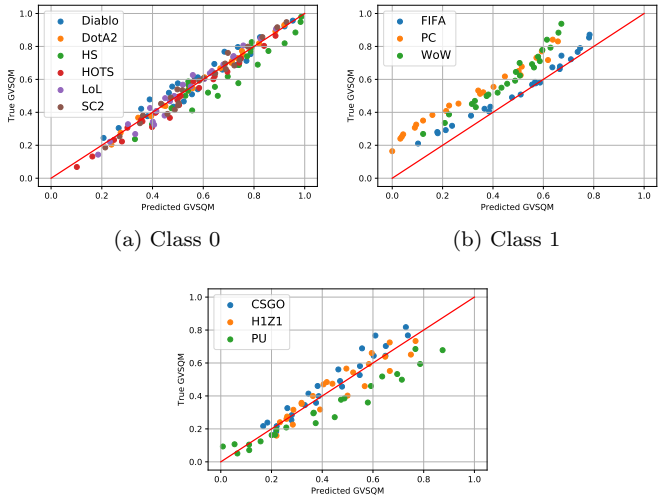


(a) (b)

Figure 15: Visualization of the shifted curves, depending on both the resolution (a) and the MMI of the particular game recording (b), here shown for the specific case of the HS game. The red-colored datapoints in Figure 15b (with lowest MMI) are the recordings used for the analysis in Section 5.

Table 10: Overview of the per-class and overall performance of the curve fitting approach.

Class	PCC	MSE
0	0.973	0.002
1	0.941	0.020
2	0.916	0.007
Overall	0.914	0.008



(a) Class 0 (b) Class 1

(c) Class 2

Figure 16: Scatterplot showing the relationship between the predicted and true values of the GVSQM metric using the curve fitting approach. The scatter plots are subdivided by class.

as well as the subdivision per class. Hereby, it should be noted that game recordings of which the data points span multiple classes are classified fully to the class that includes the most points after applying the k-means classifier. This is done to provide the model with enough data points to fit a reliable curve. As such, more reliable results will be obtained relative to a real-life case, than when a curve is fitted to only two or three data points, which would show an unrealistic decrease in performance. This is only needed within the limitations of the dataset under scrutiny, however, as real-life GVS databases can be assumed to cover enough video recordings per class such that class changes are not needed.

In Figure 16, scatter plots are provided showing the relationship between the predicted and the true GVSQM values for each of the three classes. It can be seen that class 1 shows a rather high MSE of 0.02 in comparison with its neighbouring classes, despite its 0.941 PCC. As shown in Figure 16, this is a result of a consistent under-prediction of the model, which is not the case for the other classes. Possible explanations for this behaviour might be that the SpEED-QA curves show a more complex transformation than a straightforward shift or that the proposed shift estimation is insufficient for this particular class.

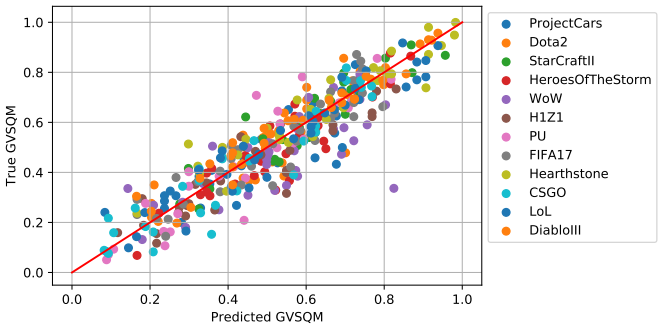
6.4. Alternatives to curve-fitting

To obtain more insight in the accuracy and scalability of the proposed framework, a comparison is provided with two alternative prediction mechanisms. A first, alternative approach makes use of a white-box DRT to capture the relationship between NR/RR-features and the GVSQM benchmark. Note that this is not exactly the same procedure as described in Section 3.2, as additional measures are taken to avoid overfitting. To this extent, a certain amount of *impurity*, *i.e.* MSE, is allowed within the trained tree. This value is optimized using 5-fold stratified CV on the dataset. Both the approach with one DRT for all data as the approach with one DRT per class are investigated. In addition, the influence of the presence/absence of SpEED-QA as a feature is researched. Table 11 shows the obtained PCC and MSE for both the "one-tree-for-all" and the "one-tree-per-class" approaches with and without the inclusion of SpEED-QA in the feature set. In Figure 17, scatter plots are provided showing the relationship between the predicted and true values of the GVSQM metric using a "one-tree-for-all" approach. Both the cases with and without the inclusion of SpEED-QA (thus creating a RR and NR approach respectively) are shown. It can be seen that a one-tree-per-class approach is showing better performance than the one-tree-for-all approach, although the gain is limited with 0.925 against 0.913 PCC for the RR case and 0.918 against 0.910 PCC for the NR case. The overall MSEs for all cases are similar. Furthermore, as one could expect, the RR approach is performing better than its NR counterpart. Again, the gain is limited, however. Somewhat surprisingly, it can be noticed that the class 2 performance is suffering the most from the exclusion of SpEED-QA from the feature set, with a 0.027 drop in PCC for the one-tree-for-all approach and a 0.018 drop in the one-tree-per-class case.

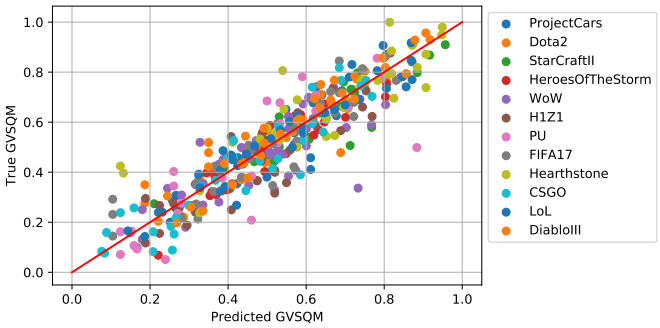
In a second alternative modelling attempt, an ANN is proposed. The activation function(s), learning rate, number of hidden neurons and number of hidden layers are chosen or optimized using a grid-search, by cross-validating on the data with 5 folds. MSE is once again chosen as the error function to be minimized. For the NR case (without SpEED-QA), a network with 2 hidden layers, 47 nodes

Table 11: Overview of the obtained PCC and MSE scores towards GVSQM for each of the three classes by applying a DRT-based approach.

	One DRT for all				One DRT per class			
	With SpEED-QA		Without SpEED-QA		With SpEED-QA		Without SpEED-QA	
Class	PCC	MSE	PCC	MSE	PCC	MSE	PCC	MSE
0	0.91	0.007	0.90	0.008	0.94	0.005	0.93	0.006
1	0.93	0.006	0.91	0.007	0.91	0.008	0.91	0.009
2	0.91	0.009	0.89	0.010	0.92	0.007	0.90	0.008



(a) With SpEED-QA

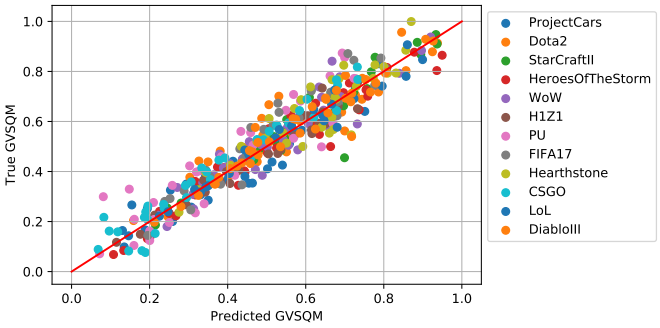


(b) Without SpEED-QA

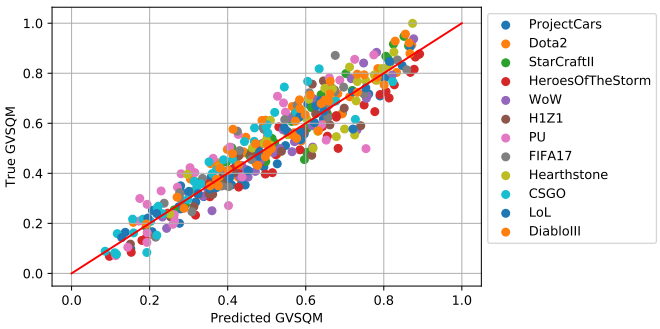
Figure 17: Scatterplot showing the relationship between the predicted and true values of the GVSQM metric using a "one-for-all" DRT-based approach. Both the model with and without the RR SpEED-QA metric are shown.

Table 12: Overview of the per-class and overall performance of the ANN model, both with and without the inclusion of the SpEED-QA RR metric.

	With SpEED-QA		Without SpEED-QA	
Class	PCC	MSE	PCC	MSE
0	0.951	0.004	0.948	0.004
1	0.952	0.004	0.959	0.003
2	0.954	0.004	0.922	0.007



(a) With SpEED-QA



(b) Without SpEED-QA

Figure 18: Scatterplot showing the relationship between the predicted and true values of the GVSQM metric using an ANN approach. Both the model with and without the RR SpEED-QA metric are shown.

per hidden layer, a 0.001 learning rate combined with a sigmoid activation function in the first layer and Rectified Linear Units (ReLUs) in the second shows to be optimal. In the RR case (with SpEED-QA), this changes towards 4 hidden layers and 29 hidden neurons per layer. The learning rate and activation functions are the same, *i.e.* one layer with a sigmoid and three with a ReLU.

The prediction results of each of the five test folds being used in the CV approach have been gathered and compared with the GVSQM benchmark. This is again done in terms of overall PCC and MSE as well as per class. Table 12 shows the overall performance for both the RR and NR model as well as per class. In Figure 18, scatter plots are provided showing the relationship between predicted and true values for both models, split per game. It can be seen that the RR model shows a rather stable performance over the three game classes, with a constant 0.004 MSE and PCC values around 0.95. The NR model shows a performance comparable to the RR model, with on overall PCC of 0.946 and MSE of 0.005 in comparison with values of 0.953 and 0.004 for the RR case. Similar to the DRT-based modelling approach, class 2 games (*i.e.* high MMI) show to suffer the most from the removal of SpEED-QA, showing a 0.032 drop in PCC and 0.003 increase in MSE in comparison with the class 2 performance in the RR case.

7. Discussion

Table 13 shows an overview of the results obtained by the multiple modelling approaches in terms of the overall PCC and MSE to GVSQM. In terms of accuracy, the ANN approach seems the most obvious choice with high

Table 13: Summary of the performance of the multiple modelling approaches in terms of the overall PCC and MSE to GVSQM

Model	PCC	MSE
ANN (RR)	0.953	0.004
ANN (NR)	0.946	0.005
One DRT for all (RR)	0.913	0.007
One DRT for all (NR)	0.910	0.008
One DRT per class (RR)	0.925	0.007
One DRT per class (NR)	0.918	0.007
Curve-fitting	0.914	0.008

PCC and low MSE values in the order of 0.95 and 0.005 respectively. In addition, the full NR approach shows only limited performance drop in comparison with the RR one, making it possible to calculate the quality estimation completely at the client-side without putting additional load on the server. The drawback, however, is the fact that an ANN approach requires a rather large set of NR-metrics to be calculated in real-time, which might be computationally unacceptable for most devices. Furthermore, the regular server-side re-training of the model on the large and ever growing video stream dataset might be a costly procedure.

The DRT approach, on the other hand, reaches acceptable overall PCCs between 0.910 and 0.925 and MSEs below 0.01, even in the NR case. However, this approach still requires the real-time, client-side calculation of a rather large NR feature set, although being much more scalable to large datasets for server-side re-training due to the models inherent logarithmic complexity (linear in worst case).

The proposed curve-fitting approach has as its most important advantage that it can be implemented with limited computational requirements of the client side, as only a single, real-time feature needs to be calculated. Despite its simplicity, it still maintains a rather high performance with a PCC of 0.914 and a MSE of 0.008 between predicted and true GVSQM. In addition, as was indicated in Table 9, an overall PCC of 0.903 and MSE of 0.012 to MOS are obtained by fitting the curve against GVSQM. Note that this RR curve-fitting approach is clearly outperforming the RR SpEED-QA metric, which only shows a -0.761 PCC to MOS. In addition, it is even outperforming the FR metrics in terms of overall correlation, whereas VMAF showed the best overall performance with a 0.864 PCC. On the downside, it has to be said that the shifting behaviour of the curves requires a rather heuristic estimation of which further research is needed to investigate its generalization towards other datasets. Moreover, as anchor points need to be transmitted and because the SpEED-QA feature seems unavoidable for class 1 games, this approach is RR by construction. However, MSEs up to 0.02 with the benchmarks are observed in worst case scenarios, which are assumed to be acceptable for most applications.

8. Conclusions

The main contribution of this scientific work is the proposition of an RR end-to-end solution for the real-time quality assessment of streamed game videos. It includes a low complexity curve-fitting approach, which is constructed based on a thorough performance analysis of objective metrics, often used for natural video quality assessment. The results show that this performance depends heavily on the game type under observation. It has been revealed that a single NR feature, *i.e.*, MMI, is sufficient to identify three different classes of games. In addition, each class has its own leading feature (JER, SpEED-QA or SC) that correlates strongly to MOS, allowing for an accurate quality assessment using a curve-fitting approach. Furthermore, a customized objective FR metric tailored to passive GVS has been created. This metric, called GVSQM, can be calculated as a weighted combination of VQM and VMAF, with the MMI acting as the weight. In addition, a comparison with other predictive models shows that the curve-fitting approach is the most promising in terms of scalability and computational complexity, but tends to show slightly lower accuracy than DRTs and ANNs. A PCC of 0.914 and MSE of 0.008 between predicted and true GVSQM is still obtained, however, resulting in a PCC of 0.903 and MSE of 0.012 when compared with a MOS benchmark. The DRT's higher accuracy comes with higher computational complexity and lower scalability and interpretability (ANN), however, and the calculation of a rather large set of real-time feature calculations on the client device (both ANN & DRT).

When it comes to further extensions to this work, it has to be said that qualitative, subjectively annotated GVS datasets are rather scarce in existing literature. Therefore, the creation of additional, possibly larger datasets could be interesting to further validate the applicability of the proposed framework as well as the objective GVSQM metric. In addition, such datasets would provide a more stable base for the further exploration of the GVS topic. Another proposed direction of research is to investigate whether or not the visual perception of an interactive gamer changes in comparison with the passive spectator. It could, for example, be possible that the active gamer lies more focus on a certain hotspot within the video, *e.g.* his/her personal avatar, while the non-interactive user keeps a more general overview of the stream. Furthermore, as Virtual Reality (VR) applications are gaining more and more attention for gaming purposes, the extension of this research towards this multimedia platform should be investigated as well. One can assume, for example, that given the hemispherical nature of VR, visual artifacts in the neighbourhood of the user's current focus point will contribute more negatively to the QoE than distortions close to the edge of the user's Field of View (FoV).

Acknowledgements

This research is part of a collaborative project between Huawei and Ghent University, funded by Huawei Technologies, China.
Maria Torres Vega is funded by the Research Foundation Flanders (FWO), grant number 12W4819N.

References

[1] K. L. Chan, Improving and Expanding Gaming Experiences based on Cloud Gaming, Ph.D. thesis, Nara Institute of Science and Technology (2018).

[2] T. Smith, M. Obrist, P. Wright, Live-Streaming Changes the (Video) Game, in: Proceedings of the 11th European Conference on Interactive TV and Video, EuroITV '13, Association for Computing Machinery, New York, NY, USA, 2013, p. 131–138. doi:10.1145/2465958.2465971.
URL https://doi.org/10.1145/2465958.2465971

[3] M. Sjöblom, J. Hamari, Why do people watch others play video games? An empirical study on the motivations of Twitch users, Computers in Human Behavior 75 (2017) 985 – 996. doi:https://doi.org/10.1016/j.chb.2016.10.019.
URL http://www.sciencedirect.com/science/article/pii/S0747563216307208

[4] M. Suznjevic, L. Skorin-Kapov, M. Matijasevic, The Impact of User, System, and Context Factors on Gaming QoE: A Case Study Involving MMORPGs, in: Proceedings of Annual Workshop on Network and Systems Support for Games, NetGames '13, IEEE Press, 2013, p. 1–6.

[5] M. Torres Vega, C. Perra, F. De Turck, A. Liotta, A Review of Predictive Quality of Experience Management in Video Streaming Services, IEEE Transactions on Broadcasting 64 (2) (2018) 432–445. doi:10.1109/TBC.2018.2822869.

[6] M. Jarschel, D. Schlosser, S. Scheuring, T. Hoßfeld, Gaming in the clouds: QoE and the users' perspective, Mathematical and Computer Modelling 57 (11) (2013) 2883 – 2894, information System Security and Performance Modeling and Simulation for Future Mobile Networks. doi:https://doi.org/10.1016/j.mcm.2011.12.014.
URL http://www.sciencedirect.com/science/article/pii/S0895717711007771

[7] I. T. Union, ITU-T RECOMMENDATION P.910 (09/99): Subjective video quality assessment methods for multimedia applications, Tech. rep., International Telecommunication Union (1999).

[8] I. T. Union, ITU-R RECOMMENDATION BT.500-14 (10/2019): Methodologies for the subjective assessment of the quality of television images, Tech. rep., International Telecommunication Union (2019).

[9] C. G. Bampis, Z. Li, A. K. Moorthy, I. Katsavounidis, A. Aaron, A. C. Bovik, Study of temporal effects on subjective video Quality of Experience, IEEE Transactions on Image Processing 26 (11) (2017) 5217–5231. doi:10.1109/TIP.2017.2729891.

[10] Y. Kang, H. Chen, L. Xie, An artificial-neural-network-based QoE estimation model for Video streaming over wireless networks, in: 2013 IEEE/CIC International Conference on Communications in China (ICCC), 2013, pp. 264–269. doi:10.1109/ICCChina.2013.6671126.

[11] M. Torres Vega, D. C. Mocanu, S. Stavrou, A. Liotta, Predictive no-reference assessment of video quality, Signal Processing: Image Communication 52 (2017) 20 – 32. doi:https://doi.org/10.1016/j.image.2016.12.001.
URL http://www.sciencedirect.com/science/article/pii/S092359651630176X

[12] M. Shahid, J. Panasiuk, G. Van Wallendael, M. Barkowsky, B. Lövström, Predicting full-reference video quality measures using HEVC bitstream-based no-reference features, in: 2015 Seventh International Workshop on Quality of Multimedia Experience (QoMEX), 2015, pp. 1–2. doi:10.1109/QoMEX.2015.7148118.

[13] N. Staelens, J. De Meulenaere, M. Claeys, G. Wallendael, W. Van den Broeck, J. De Cock, R. Van de Walle, P. Demeester, F. De Turck, Subjective quality assessment of longer duration video sequences delivered over http adaptive streaming to tablet devices, IEEE Transactions on Broadcasting 60 (2014) 707–714. doi:10.1109/TBC.2014.2359255.

[14] M. Torres Vega, D. C. Mocanu, A. Liotta, Unsupervised deep learning for real-time assessment of video streaming services, Multimedia Tools and Applications 76 (21) (2017) 22303–22327.

[15] M. Alreshoodi, J. Woods, Survey on QoE
QoS Correlation Models For Multimedia Services (2013). arXiv:1306.0221.

[16] I. Slivar, M. Suznjevic, L. Skorin-Kapov, The impact of video encoding parameters and game type on QoE for cloud gaming: A case study using the steam platform, in: 2015 Seventh International Workshop on Quality of Multimedia Experience (QoMEX), 2015, pp. 1–6. doi:10.1109/QoMEX.2015.7148144.

[17] I. Slivar, L. Skorin-Kapov, M. Suznjevic, Cloud Gaming QoE Models for Deriving Video Encoding Adaptation Strategies, in: Proceedings of the 7th International Conference on Multimedia Systems, MMSys '16, Association for Computing Machinery, New York, NY, USA, 2016, p. 18. doi:10.1145/2910017.2910602.
URL https://doi.org/10.1145/2910017.2910602

[18] I. Slivar, M. Suznjevic, L. Skorin-Kapov, Game Categorization for Deriving QoE-Driven Video Encoding Configuration Strategies for Cloud Gaming, ACM Trans. Multimedia Comput. Commun. Appl. 14 (3s) (Jun. 2018). doi:10.1145/3132041.
URL https://doi.org/10.1145/3132041

[19] N. Barman, S. Zadtootaghaj, S. Schmidt, M. G. Martini, S. Möller, GamingVideoSET: A Dataset for Gaming Video Streaming Applications, in: 2018 16th Annual Workshop on Network and Systems Support for Games (NetGames), 2018, pp. 1–6. doi:10.1109/NetGames.2018.8463362.

[20] V. Clincy, B. Wilgor, Subjective Evaluation of Latency and Packet Loss in a Cloud-Based Game, in: 2013 10th International Conference on Information Technology: New Generations, 2013, pp. 473–476. doi:10.1109/ITNG.2013.79.

[21] C.-Y. Huang, C.-H. Hsu, D.-Y. Chen, K.-T. Chen, Quantifying User Satisfaction in Mobile Cloud Games, in: Proceedings of Workshop on Mobile Video Delivery, MoViD'14, Association for Computing Machinery, New York, NY, USA, 2014, p. 1–6. doi:10.1145/2579465.2579468.
URL https://doi.org/10.1145/2579465.2579468

[22] M. Jarschel, D. Schlosser, S. Scheuring, T. Hoßfeld, An Evaluation of QoE in Cloud Gaming Based on Subjective Tests, in: 2011 Fifth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing, 2011, pp. 330–335. doi:10.1109/IMIS.2011.92.

[23] S. Wang, S. Dey, Modeling and Characterizing User Experience in a Cloud Server Based Mobile Gaming Approach, in: GLOBECOM 2009 - 2009 IEEE Global Telecommunications Conference, 2009, pp. 1–7. doi:10.1109/GLOCOM.2009.5425784.

[24] S. Zadtootaghaj, S. Schmidt, S. Möller, Modeling Gaming QoE: Towards the Impact of Frame Rate and Bit Rate on Cloud Gaming, in: 2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX), 2018, pp. 1–6. doi:10.1109/QoMEX.2018.8463416.

[25] N. Barman, M. G. Martini, S. Zadtootaghaj, S. Möller, S. Lee, A Comparative Quality Assessment Study for Gaming and Non-Gaming Videos, in: 2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX), 2018, pp. 1–6. doi:10.1109/QoMEX.2018.8463403.

[26] S. Zadtootaghaj, N. Barman, S. Schmidt, M. G. Martini, S. Möller, NR-GVQM: A No Reference Gaming Video Quality Metric, in: 2018 IEEE International Symposium on Multimedia (ISM), 2018, pp. 131–134. doi:10.1109/ISM.2018.00031.

[27] N. Barman, E. Jammeh, S. A. Ghorashi, M. G. Martini,

No-Reference Video Quality Estimation Based on Machine Learning for Passive Gaming Video Streaming Applications, IEEE Access 7 (2019) 74511–74527. doi:10.1109/ACCESS.2019.2920477.

[28] S. Göring, R. R. R. Rao, A. Raake, nofu — A Lightweight No-Reference Pixel Based Video Quality Model for Gaming Content, in: 2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX), 2019, pp. 1–6. doi:10.1109/QoMEX.2019.8743262.

[29] A. Aaron, Z. Li, M. Manohara, J. Y. Lin, E. C. Wu, C. . J. Kuo, Challenges in cloud based ingest and encoding for high quality streaming media, in: 2015 IEEE International Conference on Image Processing (ICIP), 2015, pp. 1732–1736. doi:10.1109/ICIP.2015.7351097.

[30] M. Fiedler, T. Hossfeld, P. Tran-Gia, A generic quantitative relationship between Quality of Experience and Quality of Service, IEEE Network 24 (2) (2010) 36–41. doi:10.1109/MNET.2010.5430142.

[31] E. Alpaydin, Introduction to Machine Learning, 3rd Edition, MIT Press, Cambridge, Massachusetts, USA / London, England, 2014.

[32] A. Liotta, D. C. Mocanu, V. Menkovski, L. Cagnetta, G. Exarchakos, Instantaneous Video Quality Assessment for Lightweight Devices, in: Proceedings of International Conference on Advances in Mobile Computing & Multimedia, MoMM '13, Association for Computing Machinery, New York, NY, USA, 2013, p. 525–531. doi:10.1145/2536853.2536903.
URL https://doi.org/10.1145/2536853.2536903

[33] M. G. Choi, J. H. Jung, J. W. Jeon, No-reference image quality assessment using blur and noise, International Journal of Computer Science and Engineering 3 (2) (2009) 76–80.

[34] C. Perra, A low computational complexity blockiness estimation based on spatial analysis, in: 2014 22nd Telecommunications Forum Telfor (TELFOR), 2014, pp. 1130–1133. doi:10.1109/TELFOR.2014.7034606.

[35] P. Paudyal, F. Battisti, M. Carli, Impact of video content and transmission impairments on Quality of Experience, Multimedia Tools and Applications 75 (23) (2016) 16461–16485. doi:10.1007/s11042-015-3214-0.
URL https://doi.org/10.1007/s11042-015-3214-0

[36] S. Borer, A model of jerkiness for temporal impairments in video transmission, in: 2010 Second International Workshop on Quality of Multimedia Experience (QoMEX), 2010, pp. 218–223. doi:10.1109/QOMEX.2010.5516155.

[37] S. Zadtootaghaj, S. Schmidt, N. Barman, S. Möller, M. G. Martini, A Classification of Video Games based on Game Characteristics linked to Video Coding Complexity, in: 2018 16th Annual Workshop on Network and Systems Support for Games (NetGames), 2018, pp. 1–6. doi:10.1109/NetGames.2018.8463434.

[38] FFmpeg, www.ffmpeg.org (2019).

[39] M. H. Pinson, S. Wolf, A new standardized method for objectively measuring video quality, IEEE Transactions on Broadcasting 50 (3) (2004) 312–322. doi:10.1109/TBC.2004.834028.

[40] Video Quality Metric (VQM) Software, Institute for Telecommunication Sciences, https://www.its.bldrdoc.gov/resources/video-quality-research/software.aspx (2019).